ELSEVIER Stochastic Processes and their Applications 73 (1998) 101–118

stochastic processes and their applications

Markov-achievable payoffs for finite-horizon decision models

Victor Pestien [a,*], Xiaobo Wang [b]

[a] *Department of Mathematics and Computer Science, University of Miami, Coral Gables, FL 33124, USA*

[b] *Jostens Learning Corporation, 9820 Pacific Heights Blvd. San Diego, CA 92121, USA*

Received 24 February 1997; received in revised form 19 June 1997

Abstract

Consider the class of n-stage decision models with state space S, action space A, and payoff function $g : (S \times A)^n \times S \to \mathbb{R}$. The function g is *Markov-achievable* if for any possible set of available randomized actions and all transition laws, each plan has a corresponding Markov plan whose value is at least as good. A condition on g, called the "non-forking linear sections property", is necessary and sufficient for g to be Markov achievable. If g satisfies the slightly stronger "general linear sections property", then g can be written as a sum of products of certain simple neighboring-stage payoffs.

AMS classifications: primary 90C40; secondary 60G40, 60K15, 90C39

Keywords: Markov decision model; Payoff function; Markov plan

1. Introduction and framework

A fundamental question in the study of Markov decision models asks under what conditions it is optimal, or at least nearly optimal, to use a Markov selection rule at each stage. It is well known that this question has different answers depending on what optimality criterion (payoff function) is being used. Thus, it is natural to ask which payoff functions g will always lead to the existence of good Markov selection rules at each stage of a finite-stage decision problem.

Here is an informal description of the results to be obtained in this note (more precise definitions and statements will follow later): Let S and A be the state and action spaces. We say that a real-valued payoff function g with domain $(S \times A)^n \times S$ is Markov-achievable if for every possible set of available randomized actions and every possible transition law p, and each plan π, there is a Markov plan $\hat{\pi}$ such that the average of g under $\hat{\pi}$ is at least that under π. Assuming that the action space has at least three elements, we establish a condition on g, called the "non-forking linear sections property", which is necessary and sufficient for g to be Markov-achievable.

* Corresponding author. E-mail: vcp@cs.cs.miami.edu.

0304-4149/98/$19.00
PII S0304-4149(97)00095-1

We also give a slightly-stronger condition, the "general linear sections property", and show that each g having this property can be written as a sum of certain products involving functions of states and actions which are neighboring in time.

We now begin by giving a formal definition of the model.

Definition 1.1. Suppose n is a positive integer. An n-stage decision model is a five-tuple $(S, A, \mathcal{Q}, p, g)$, where

(1) the *state space* S is a non-empty, countable set;
(2) the *action space* A is a non-empty, countable set;
(3) the *randomized action map* $\mathcal{Q}$ assigns to each s in S and each k $(1 \leqslant k \leqslant n)$ a non-empty countable subset $\mathcal{Q}(s,k)$ of the set of probability measures on the subsets of A. The set $\mathcal{Q}(s,k)$ is the set of *available randomized actions* at state s and time k;
(4) the *transition law* p associates with each s, each k, and each a in A a probability measure $p(\cdot|s,k,a)$ on the subsets of S;
(5) the *payoff function* g is a bounded mapping from the set $(S \times A)^n \times S$ of histories of length n to the set $\mathbb{R}$ of real numbers.

At time 0, the system begins at state s_0 which is chosen from S according to some initial distribution p_0. Then action a_1 is chosen using a measure in $\mathcal{Q}(s_0, 1)$, and the system moves to state s_1, where s_1 is selected according to the transition law $p(\cdot|s_0, 1, a_1)$. Next, action a_2, chosen with a measure from $\mathcal{Q}(s_1, 2)$, causes a move to s_2, where s_2 is selected according to $p(\cdot|s_1, 2, a_2)$. The procedure continues until time n, when the state s_n is reached. The payoff received for this procedure is $g(s_0 a_1 s_1 a_2 \cdots s_{n-1} a_n s_n)$.

The collections of *histories* are denoted by $H_0 = S$ and by

$$H_k := (S \times A)^k \times S \qquad (1 \leqslant k \leqslant n),$$

so that the payoff function g has domain H_n. The sequence of rules for selecting the actions in the decision process described above is called a *plan*. Thus, a plan $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ is a sequence of *selection rules*

$$\pi_k : H_{k-1} \to \mathcal{Q}(s,k).$$

If a plan π is paired with an initial distribution p_0 on the subsets of S, then a probability measure P_{π,p_0} is induced on H_n so that for each $s_0 a_1 s_1 a_2 \cdots s_{n-1} a_n s_n$ in H_n,

$$P_{\pi,p_0}(\{s_0 a_1 s_1 a_2 \cdots s_n\}) = p_0(\{s_0\}) \prod_{k=1}^{n} p(s_k|s_{k-1}, k, a_k) \pi_k(s_0 a_1 \cdots s_{k-1})(\{a_k\}).$$

For $1 \leqslant k \leqslant n$, a selection rule π_k is *Markov* if for each $q = s_0 a_1 s_1 \cdots a_k s_k$ in H_k and each $q' = s'_0 a'_1 s'_1 \cdots a'_k s'_k$ in H_k, we have $\pi_k(q) = \pi_k(q')$ whenever $s_k = s'_k$. A plan π is *time-*$[k,n]$*Markov* if the selection rules $\pi_k, \pi_{k+1}, \ldots, \pi_n$ are Markov. A plan π is a *Markov plan* if it is time-$[1,n]$ Markov. Notice that for a Markov plan, the action which π takes at state s and time j depends only on s and j and not on any previous states visited or previous actions taken.

Definition 1.2. The function $g : H_n \to \mathbb{R}$ is *Markov-achievable* if for every mapping $\mathcal{Q}$ and every transition law p which make $(S, A, \mathcal{Q}, p, g)$ an n-stage decision model, and for each initial distribution p_0 and each plan π, there is a Markov plan $\hat{\pi}$ such that

$$\int g \, dP_{\hat{\pi}} \geqslant \int g \, dP_{\pi}. \tag{1.1}$$

In Eq. (1.1) we have written the measures $P_{\hat{\pi},p_0}$ and P_{π,p_0} without the qualifier "p_0". In what follows, we will continue to suppress the "p_0" whenever there is no ambiguity.

Dynkin and Yushkevich (1979) give a systematic treatment of the theory of Markov decision processes, both on finite and infinite time intervals. They establish the sufficiency of Markov plans for a large class of models. Feinberg (1982) demonstrates the sufficiency of Markov plans for models where payoff functions satisfy certain very general criteria. Larson and Casti (1978) illustrate some dynamic programming reward criteria which have the "Markovian property", a notion similar to our "Markov-achievable payoff function". Hill and Pestien (1987) identify a collection of payoff functions which turn out to be Markov-achievable.

Pioneering work on finite-stage Markov decision models was done by Bellman, (1957). Fundamental results for related dynamic programming problems were established by Blackwell (1965). The work by Hinderer (1970) provides a careful history of the development of decision processes and uses a broad model in which payoffs may depend on the entire history of the process. Schäl and Sudderth (1987) prove the uniform adequacy of Markov plans for a wide-ranging model which includes the usual dynamic programming payoff as well as others. The text by White (1993) gives excellent insight and a wealth of examples and exercises about Markov decision models.

2. Examples and the general linear sections property

We begin this section with two very simple examples of functions which fail to be Markov-achievable. The reasoning underlying these examples will be generalized as part of a proof in the latter half of Section (4).

Example 2.1. Let S have only one element, denoted by "†", let $A = \{a, b\}$, and let $n = 2$. Define g on H_2 by

$$g(\dagger a \dagger a\dagger) = g(\dagger b \dagger b\dagger) = 1$$

and

$$g(\dagger a \dagger b\dagger) = g(\dagger b \dagger a\dagger) = 0.$$

(So the payoff is 1 if the actions used at times 1 and 2 are the same, and the payoff is 0 otherwise.) Let $Q(\dagger, 1)$ contain only the measure that gives probability $\frac{1}{2}$ to $\{a\}$ and $\frac{1}{2}$ to $\{b\}$. Let $Q(\dagger, 2)$ contain all probability measures on subsets of A. If the plan $\hat{\pi}$ is Markov, then the measures $\hat{\pi}(\dagger a\dagger)$ and $\hat{\pi}(\dagger b\dagger)$ must be the same, and therefore

$$\int g \, dP_{\hat{\pi}} = \frac{1}{2}.$$

However, suppose π is a plan involving the non-Markov selection rule π_2 for which

$$\pi_2(\dagger a\dagger)(\{a\}) = \pi_2(\dagger b\dagger)(\{b\}) = 1.$$

Since P_π lives on the two histories $\dagger a \dagger a\dagger$ and $\dagger b \dagger b\dagger$, we have

$$\int g \, dP_\pi = 1.$$

Thus, g is not Markov-achievable.

Example 2.2. Let $S = \{\dagger\}$, let $A = \{a, b, c\}$, and let $n = 2$. Define g on H_2 by

$$g(\dagger a \dagger a\dagger) = g(\dagger b \dagger a\dagger) = g(\dagger b \dagger b\dagger\} = 1$$

and $g(h) = 0$ otherwise. Again let $Q(\dagger, 1)$ contain only the measure that gives probability $\frac{1}{2}$ to $\{a\}$ and $\frac{1}{2}$ to $\{b\}$. This time let $Q(\dagger, 2) = \{\mu, \nu\}$, where $\nu(\{b\}) = 1$ and

$$\mu(\{a\}) = \mu(\{c\}) = \tfrac{1}{2}.$$

It is easy to see that for any Markov plan $\hat{\pi}$,

$$\int g \, dP_{\hat{\pi}} = \tfrac{1}{2}.$$

On the other hand, if π involves any non-Markov selection rule π_2 for which

$$\pi_2(\dagger b\dagger) = \nu \quad \text{and} \quad \pi_2(\dagger a\dagger) = \mu,$$

then

$$P_\pi(\dagger a \dagger a\dagger) = P_\pi(\dagger a \dagger c\dagger) = 1/4 \quad \text{and} \quad P_\pi(\dagger b \dagger b\dagger) = \tfrac{1}{2}.$$

Hence,

$$\int g \, dP_\pi = \tfrac{3}{4},$$

and g is not Markov-achievable.

Definition 2.3. A collection of points (not necessarily distinct) in $\mathbb{R}^2$ is *strongly collinear* if all of the points coincide or if all of the points lie on the same straight line and that line either is horizontal, is vertical, or has positive slope. (Thus, lines of negative slope are disallowed.)

Lemma 2.4. *If three points* $(r_1, s_1), (r_2, s_2)$, *and* (r_3, s_3) *in* $\mathbb{R}^2$ *are not strongly collinear, then there exists a real number* ξ *such that* $0 < \xi \leqslant 1$ *and for some ordering* (i, j, k) *of* $(1, 2, 3)$,

$$\frac{\xi s_i + (1-\xi)s_k - s_j}{\xi r_i + (1-\xi)r_k - r_j} < 0. \tag{2.1}$$

Proof. If the three non-strongly collinear points are distinct, and if they can be arranged so that for some ordering (i,j,k) of $(1,2,3)$

$$r_i \leqslant r_j \leqslant r_k \qquad \text{and} \qquad s_i \leqslant s_j \leqslant s_k,$$

then for some ξ in $(0,1)$, the slope of the line through

$$(\xi r_i + (1-\xi) r_k, \xi s_i + (1-\xi) s_k) \qquad \text{and} \qquad (r_j, s_j)$$

is negative and so Eq. (2.1) holds. If the points cannot be so arranged, or if two of the points coincide, then a line through some pair of points has negative slope, and we can take $\xi = 1$ in Eq. (2.1). □

Definition 2.5. The payoff function $g : H_n \to \mathbb{R}$ has the *general linear sections property* (*GLSP*) if for each s in S, each k $(1 \leqslant k \leqslant n-1)$, each q' in $(S \times A)^k$ and each q'' in $(S \times A)^k$, the set

$$\{(g(q'sw), g(q''sw)) : w \in (A \times S)^{n-k}\} \tag{2.2}$$

is strongly collinear.

To see how the definition of GLSP fits Examples 2.1 and 2.2, let $s = \dagger$, let $q' = \dagger a$, and let $q'' = \dagger b$. Then in Example 2.1

$$\{(g(\dagger a \dagger w), g(\dagger b \dagger w)) : w \in A \times S\} = \{(1,0),(0,1)\},$$

a set which is not strongly collinear, while in Example 2.2,

$$\{(g(\dagger a \dagger w), g(\dagger b \dagger w)) : w \in A \times S\} = \{(1,1),(0,1),(0,0)\},$$

another non-strongly-collinear set. Thus, neither of the functions g has the GLSP.

In the third example of this section, we describe a classical payoff function which is known to be Markov-achievable:

Example 2.6. As in the classical dynamic programming setup, let

$$g(s_0 a_1 s_1 \cdots a_n s_n) = \sum_{i=1}^{n} u(s_{i-1}, a_i, s_i),$$

where the "utility function" $u : S \times A \times S \to \mathbb{R}$ is bounded. If $q' \in (S \times A)^k$, $q'' \in (S \times A)^k$, and $s^* \in S$, then all points in the set

$$\{(g(q's^*w), g(q''s^*w)) : w \in (A \times S)^{n-k}\}$$

lie on the same straight line of slope 1. To see this, notice that if $w' = a'_{k+1}s'_{k+1} \cdots a'_n$ and $w'' = a''_{k+1}s''_{k+1} \cdots a''_n$, then

$$g(q''s^*w'') - g(q''s^*w') = \sum_{i=k+1}^{n} u(s''_{i-1}, a''_i, s''_i) - \sum_{i=k+1}^{n} u(s'_{i-1}, a'_i, s'_i)$$

$$= g(q's^*w'') - g(q's^*w'),$$

where s''_k and s'_k both denote s^*. Therefore, g has the GLSP.

It will follow from later results that if g has the GLSP, then g must be Markov-achievable. However, the converse turns out not to be true. Thus, in order to obtain a condition equivalent to Markov achievability, we must formulate a stricter type of linear sections property.

3. Non-forking paths and the NFLSP

The chief theorem, to be stated later in this section, will use the notion of "non-forking" paths and the non-forking linear sections property.

Definition 3.1. Let $w \in (A \times S)^m$ and $\hat{w} \in (A \times S)^m$, where $w = a_1 s_1 \cdots s_m$ and $\hat{w} = \hat{a}_1 \hat{s}_1 \cdots \hat{s}_m$. Then w and $\hat{w}$ are *non-forking paths* if, whenever $s_\ell = \hat{s}_\ell$ for some ℓ $(1 \leqslant \ell \leqslant m)$, we also have $a_j = \hat{a}_j$ and $s_j = \hat{s}_j$ for all j such that $\ell < j \leqslant m$.

Thus, w and $\hat{w}$ are non-forking paths if, having agreed at some state, they agree at all actions and states that follow. Notice that even if w and $\hat{w}$ never agree, they are still called *non-forking*. Three paths $w, \hat{w}$, and $\hat{\hat{w}}$ are *non-forking* if each pair w and $\hat{w}$,w and $\hat{\hat{w}}$, and $\hat{w}$ and $\hat{\hat{w}}$ is non-forking.

Definition 3.2. The payoff function $g : H_n \to \mathbb{R}$ has the *non-forking linear sections property* (*NFLSP*) if for each s in S, each k $(1 \leqslant k \leqslant n-1)$, each q' in $(S \times A)^k$ and each q'' in $(S \times A)^k$, and for any three non-forking paths w^1, w^2, and w^3 in $(A \times S)^{n-k}$ whose initial actions are all distinct, the points

$$(g(q' s w^i), g(q'' s w^i)), \quad i = 1, 2, 3$$

are strongly collinear.

The following proposition is obvious from Definitions 2.5 and 3.2:

Proposition 3.3. *If $g : H_n \to \mathbb{R}$ has the general linear sections property, then g has the non-forking linear sections property.*

Here are some examples illustrating the NFLSP:

Example 3.4. Let $S = \{1, 2, \ldots\}$ and suppose A contains at least the three distinct actions a, b, and c. For $s_0 a_1 s_1 \cdots a_n s_n$ in H_n, let

$$g(s_0 a_1 s_1 \cdots a_n s_n) = \max\{s_0, s_1, \ldots, s_n\}.$$

Thus, the payoff g for any path is the maximum value among all states visited, regardless of actions taken. Then g does not have the NFLSP. To show this, in the notation of Definition 3.2 let $s = 1$ and suppose q' visits only the "1" state and q'' visits only the "2" state, while w^1 always uses action a and visits only "1", w^2 uses b and visits only "2", and w^3 uses c and visits only "3". Then w^1 and w^2 and w^3 are non-forking,

and the following set is not strongly collinear:

$$\{(g(q'sw^1), g(q''sw^1)), (g(q'sw^2), g(q''sw^2)), (g(q'sw^3), g(q''sw^3))\}$$
$$= \{(1,2),(2,2),(3,3)\}.$$

Thus, g does not have the NFLSP.

Example 3.5. Let $S = \{\dagger\}$, $A = \{a, b, c\}$, and $n = 3$. Define g on H_3 by

$$g(h) = \begin{cases} 1 \text{ if } h \text{ uses the action } a \text{ at least twice,} \\ 0 \text{ otherwise.} \end{cases}$$

It is easy to check that the NFLSP holds. Also, it is intuitively clear that g is Markov-achievable, because for any randomized action map $\mathcal{Q}$, a good Markov plan can be obtained by always choosing (at the single state $\dagger$) those randomized actions which make the probability of $\{a\}$ large. But g does not have the GLSP because, in the notation of Definition 2.5, if $q' = \dagger a\dagger$, $q'' = \dagger b\dagger$, $w^1 = a \dagger a\dagger$, $w^2 = a \dagger b\dagger$, and $w^3 = b \dagger b\dagger$, then

$$\{(g(q'sw^1), g(q''sw^1)), (g(q'sw^2), g(q''sw^2)), (g(q'sw^3), g(q''sw^3))\}$$
$$= \{(1,1),(1,0),(0,0)\}, \tag{3.1}$$

a set which is not strongly collinear.

However, notice that we can modify the previous example slightly to obtain a g which does *not* satisfy the NFLSP:

Example 3.6. In Example 3.5, suppose instead that S has at least three elements s, u, and v, but that A, g, and n remain the same. In the notation of Definition 3.2, let $q' = sas, q'' = sbs, w^1 = asas, w^2 = buas$, and $w^1 = cvbs$. Then the paths w^1, w^2, and w^3 are non-forking, but the set described in Eq. (3.1) is not strongly collinear. Thus, g does not have the NFLSP.

We are now prepared to state the main theorem of this paper. The proof will be given in the next section.

Theorem 3.7. *Suppose A and S are countable sets with $|A| \geqslant 3$ and S non-empty, and suppose*

$$g : (S \times A)^n \times S \to \mathbb{R}.$$

Then g has the non-forking linear sections property if and only if g is Markov-achievable.

In Pestien and Wang (1993), somewhat-related questions were studied for a finite-horizon payoff function which depended only on the states visited, not on the actions taken. A "linear sections property" was formulated there and was shown to be sufficient for "Markov adequacy". (A function g on S^n has the Markov adequacy property if every strategy has a corresponding Markov strategy which gives g the same integral as the original strategy.) The same paper also showed that the linear sections property

was necessary for Markov adequacy, but unfortunately only under the very restrictive hypothesis that g be permutation invariant.

The formulation in Pestien and Wang (1993) follows closely the gambling-theoretic framework of Dubins and Savage (1976), while the formulation in the current note is closer to the dynamic programming setup of Blackwell (1965) and others. It is possible to define a correspondence between gambling problems and dynamic programming problems, but there is no nice pairing between Markov plans in the one framework and those in the other. An extensive discussion of gambling problems and of their relations to dynamic programming is given in the recent book by Maitra and Sudderth (1996, Ch. 3).

The following two technical lemmas give necessary conditions for a payoff function g to have the non-forking linear sections property. To introduce notation for the lemmas, if $q \in (S \times A)^k$, $s \in S$, and $a \in A$, let g_{qsa} denote the function defined on H_{n-k-1} by

$$g_{qsa}(h) = g(qsah).$$

For abbreviation, let

$$\{qs-\} = \{qsw\colon w \in (A \times S)^{n-k}\}$$

and

$$\{qsa-\} = \{qsav\colon v \in H_{n-k-1}\}.$$

We say qs (resp. qsa) *is feasible under a plan* π if

$$P_\pi(\{qs-\}) > 0 \qquad (\text{resp. } P_\pi(\{qsa-\}) > 0).$$

Lemma 3.8. *Let $g : H_n \to \mathbb{R}$ be a payoff function and let $1 \leqslant k \leqslant n-1$. Suppose there exist q in $(S \times A)^k$, $\hat{q}$ in $(S \times A)^k$, s in S, distinct actions a^1, a^2, and a^3 in A, and there exists an initial distribution p_0 and a time-$[k+2, n]$ Markov plan π such that the points Q_1, Q_2, and Q_3, are not strongly collinear, where*

$$Q_i = \left(\int_{H_{n-k-1}} g_{qsa^i} \, dP_\pi, \int_{H_{n-k-1}} g_{\hat{q}sa^i} \, dP_\pi \right), \qquad i = 1,2,3$$

and qsa^i and $\hat{q}sa^i$ are feasible under π. Then the function g does not have the non-forking linear sections property.

Note: In the extreme case where $k = n-1$, we regard *any* plan π as being time-$[n+1, n]$ Markov.

Proof. For each i, define $\mu_0^i(\cdot) = p(\cdot | s, k+1, a^i)$. Since Q_1, Q_2, and Q_3 are averages of a fixed set of points with respect to the probability measures μ_0^1, μ_0^2, and μ_0^3, respectively, we can find states s^1, s^2, and s^3 such that if

$$\tilde{Q}_i = \left(\int_{(A\times S)^{n-k-1}} g_{qsa^is^i} \, dP_\pi, \int_{(A\times S)^{n-k-1}} g_{\hat{q}sa^is^i} \, dP_\pi \right), \quad i = 1,2,3$$

and qsa^is^i and $\hat{q}sa^is^i$ are feasible under π, then $\tilde{Q}_1, \tilde{Q}_2$, and $\tilde{Q}_3$ are not strongly collinear.

If $k = n-1$, then a^1s^1, a^2s^2, and a^3s^3 are the desired non-forking paths. Otherwise, there are now three cases to consider:

Case 1: $s^1 = s^2 = s^3$ (call the common value s^*). Since the plan π is Markov and always begins at the same state, each point $\tilde{Q}_1$, $\tilde{Q}_2$, and $\tilde{Q}_3$ is an average with respect to the same probability measure induced by the plan π. Therefore, there exists v in $(A \times S)^{n-k-1}$ such that qsa^is^*v and $\hat{q}sa^is^*v$ are feasible under π and such that the points

$$Q_i^* = (g(qsa^is^*v), g(\hat{q}sa^is^*v)), \quad i = 1,2,3$$

are not strongly collinear. Then $w^1 = a^1s^*v$ and $w^2 = a^2s^*v$ and $w^3 = a^3s^*v$ are the desired non-forking paths.

Case 2: Exactly two of s^1, s^2, and s^3 are equal (say $s^* := s^1 = s^2$). Then we can find states $\bar{s}$ and $\bar{\bar{s}}$ and non-strongly collinear points $Q_1, \tilde{\tilde{Q}}_2$, and $\tilde{\tilde{Q}}_3$, where

$$\tilde{\tilde{Q}}_i = \left(\int_{(A\times S)^{n-k-2}} g_{qsa^is^*\bar{a}\bar{s}} \, dP_\pi, \int_{(A\times S)^{n-k-2}} g_{\hat{q}sa^is^*\bar{a}\bar{s}} \, dP_\pi \right), \qquad i = 1,2$$

$qsa^is^*\bar{a}\bar{s}$ and $\hat{q}sa^is^*\bar{a}\bar{s}$ are feasible under π, and

$$\tilde{\tilde{Q}}_3 = \left(\int_{(A\times S)^{n-k-2}} g_{qsa^3s^3\bar{\bar{a}}\bar{\bar{s}}} \, dP_\pi, \int_{(A\times S)^{n-k-2}} g_{\hat{q}sa^3s^3\bar{\bar{a}}\bar{\bar{s}}} \, dP_\pi \right),$$

where $qsa^3s^3\bar{\bar{a}}\bar{\bar{s}}$ and $\hat{q}sa^3s^3\bar{\bar{a}}\bar{\bar{s}}$ are feasible under π.

If $\bar{s} = \bar{\bar{s}}$, then we can find v' in $(A \times S)^{n-k-1}$ such that $qsa^is^*\ \bar{a}\bar{s}v'$ and $\hat{q}sa^is^*\bar{a}\bar{s}v'$ $(i = 1,2)$ and $qsa^3s^3\bar{\bar{a}}\bar{\bar{s}}v'$ and $\hat{q}sa^3s^3\bar{\bar{a}}\bar{\bar{s}}v'$ are feasible under π and such that the points

$$(g(qsa^1s^*\bar{a}\bar{s}v'), g(\hat{q}sa^1s^*\bar{a}\bar{s}v')), \quad (g(qsa^2s^*\bar{a}\bar{s}v'), g(\hat{q}sa^2s^*\bar{a}\bar{s}v')),$$

and

$$(g(qsa^3s^3\,\bar{\bar{a}}\bar{\bar{s}}v'), g(\hat{q}sa^3s^3\,\bar{\bar{a}}\bar{\bar{s}}v'))$$

are not strongly collinear.

If $\bar{s} \neq \bar{\bar{s}}$, then continue the procedure until agreeing states are found and thus non-forking paths are identified. If no agreeing states are ever found, then the appropriate paths remain non-forking.

Case 3: All of s^1, s^2, and s^3 are different. Then we can find non-strongly collinear points $\tilde{\tilde{Q}}_1, \tilde{\tilde{Q}}_2$, and $\tilde{\tilde{Q}}_3$, where

$$\tilde{\tilde{Q}}_i = \left(\int_{(A\times S)^{n-k-2}} g_{qsa^is^i\bar{a}^i\bar{s}^i} \, dP_\pi, \int_{(A\times S)^{n-k-2}} g_{\hat{q}sa^is^i\bar{a}^i\bar{s}^i} \, dP_\pi \right), \qquad i = 1,2,3$$

and $qsa^is^i\bar{a}^i\bar{s}^i$ and $\hat{q}sa^is^i\bar{a}^i\bar{s}^i$ are feasible under π. Then either all of $\bar{s}^1, \bar{s}^2$, and $\bar{s}^3$ are the same, or exactly two are the same, or all are different. In any of these subcases, we can proceed as above to eventually identify the appropriate non-forking paths. □

Lemma 3.9. *Suppose* $g : H_n \to \mathbb{R}$ *has the non-forking linear sections property. Then for each* k $(1 \leqslant k \leqslant n-1)$ *and each time-*$[n-k+2, n]$ *Markov plan* π, *there exist*

bounded mappings $\alpha_{n-k} : H_{n-k} \to [0,\infty)$, $\beta_{n-k} : H_{n-k} \to \mathbb{R}$, *and* $\gamma_{n-k} : H_k \to \mathbb{R}$ *such that for each* q *in* $(S \times A)^{n-k}$, *each* s *in* S *and each* a *in* A,

$$\int_{H_{k-1}} g_{qsa}\, dP_\pi = \alpha_{n-k}(qs) \int_{H_{k-1}} (\gamma_{n-k})_{sa}\, dP_\pi + \beta_{n-k}(qs).$$

Proof. Use the contrapositive of the previous lemma and put the straight line in slope-intercept form. □

4. The NFLSP and Markov-achievability

In this section we develop the proof of Theorem 3.7. First, we state an easy lemma which will be used below:

Lemma 4.1. *Let* $\{r_1, r_2, \ldots\}$ *be a bounded sequence of real numbers and* $\{c_1, c_2, \ldots\}$ *be a sequence of non-negative real numbers with* $\sum_{i=1}^{\infty} c_i < \infty$. *Then there exists a real number* r *in* $\{r_1, r_2, \ldots\}$ *such that*

$$r \sum_{i=1}^{\infty} c_i \geqslant \sum_{i=1}^{\infty} c_i r_i.$$

Proof that NFLSP implies Markov achievability. Let $(S, A, \mathcal{Q}, p, g)$ be an n-stage decision model with initial distribution p_0. Given the plan π, we will refine the standard algorithm of backward induction to build a sequence of plans $\pi^{(0)}, \pi^{(1)}, \ldots, \pi^{(n-1)}$ such that for each j, $\pi^{(j)}$ is time-$[n-j+1, n]$ Markov and

$$\int g\, dP_{\pi^{(j)}} \geqslant \int g\, dP_\pi. \tag{4.1}$$

Then, after constructing the sequence, we will let $\hat{\pi} = \pi^{(n-1)}$, which is the desired Markov plan.

To begin (step 0), let $\pi^{(0)}$ be the plan π. As in the note following Lemma 3.8, $\pi^{(0)}$ is regarded as being time-$[n+1, n]$ Markov. For the mth step $(1 \leqslant m \leqslant n-1)$, suppose from step $m-1$ we have a time-$[n-m+2, n]$ Markov plan $\pi^{(m-1)}$ such that

$$\int g\, dP_{\pi^{(m-1)}} \geqslant \int g\, dP_\pi.$$

If g has the non-forking linear sections property, then by Lemma 3.9, we have mappings $\alpha_{n-m} : H_{n-m} \to [0,\infty)$, $\beta_{n-m} : H_{n-m} \to \mathbb{R}$, and $\gamma_{n-m} : H_m \to \mathbb{R}$ such that for each q in $(S \times A)^{n-m}$, each s in S, and each a in A such that qsa is feasible under $\pi^{(m-1)}$,

$$\int_{H_{m-1}} g_{qsa}\, dP_{\pi^{(m-1)}} = \alpha_{n-m}(qs) \int_{H_{m-1}} (\gamma_{n-m})_{sa}\, dP_{\pi^{(m-1)}} + \beta_{n-m}(qs). \tag{4.2}$$

Fix s in S. If qs is feasible under $\pi^{(m-1)}$, let

$$r_{qs}^{(m-1)} = \sum_{w \in (A \times S)^m} \frac{P_{\pi^{(m-1)}}(\{qsw\})}{P_{\pi^{(m-1)}}(\{qs-\})} \gamma_{n-m}(sw). \tag{4.3}$$

We can think of $r_{qs}^{(m-1)}$ as the conditional expected value of $\gamma_{n-m}(s-)$ under the plan $\pi^{(m-1)}$, given that the history qs has occurred. We wish to make $\gamma_{n-m}(s-)$ as large as possible when at state s at time $n-m$. To do so, we will use a randomized action which had arisen under plan $\pi^{(m-1)}$ from a history q^*s which made $r_{q^*s}^{(m-1)}$ large.

Now, for any q in $(S \times A)^{n-m}$, let

$$c_{qs}^{(m-1)} = P_{\pi^{(m-1)}}(\{qs-\})\alpha_{n-m}(qs)$$

and notice that each $c_{qs}^{(m-1)}$ is non-negative,

$$\sum_{q \in (S \times A)^{n-m}} c_{qs}^{(m-1)} < \infty,$$

and the set $\{r_{qs}^{(m-1)}: q \in (S \times A)^{n-m}\}$ is bounded. Then by Lemma 4.1, there exists q^* in $(S \times A)^{n-m}$ such that q^*s is feasible for $\pi^{(m-1)}$ and

$$r_{q^*s}^{(m-1)} \sum_{q \in (S \times A)^{n-m}} c_{qs}^{(m-1)} \geqslant \sum_{q \in (S \times A)^{n-m}} [c_{qs}^{(m-1)} r_{qs}^{(m-1)}]. \tag{4.4}$$

Thus, for each s in S, we have identified a particular history q^*s which leads to a large payoff under $\pi^{(m-1)}$. Hence, we will arrange the Markov selection rule $\pi_{n-m+1}^{(m)}$ so that $\pi_{n-m+1}^{(m)}(qs)$ always uses $\pi_{n-m+1}^{(m-1)}(q^*s)$, regardless of q. That is, define the time-$[n-m+1, n]$ Markov plan $\pi^{(m)}$ as follows:

$$\pi_{n-m+1}^{(m)}(qs) = \pi_{n-m+1}^{(m-1)}(q^*s) \qquad \text{if} \quad q \in (S \times A)^{n-m} \tag{4.5}$$

and

$$\pi_k^{(m)}(h) = \pi_k^{(m-1)}(h) \qquad \text{if} \quad h \in H_{k-1} \text{ and } k \neq n-m+1. \tag{4.6}$$

By Eq. (4.6), for each q in $(S \times A)^{n-m}$,

$$P_{\pi^{(m)}}(\{qs-\}) = P_{\pi^{(m-1)}}(\{qs-\}). \tag{4.7}$$

Thus, we have, for each s in S,

$$\sum_{q \in (S \times A)^{n-m}} \sum_{w \in (A \times S)^m} \alpha_{n-m}(qs)\gamma_{n-m}(sw) P_{\pi^{(m)}}(\{qsw\})$$

$$= \sum_{q \in (S \times A)^{n-m}} P_{\pi^{(m)}}(\{qs-\}) \sum_{w \in (A \times S)^m} \frac{P_{\pi^{(m)}}(\{qsw\})}{P_{\pi^{(m)}}(\{qs-\})} \alpha_{n-m}(qs)\gamma_{n-m}(sw)$$

$$= \sum_{q \in (S \times A)^{n-m}} P_{\pi^{(m-1)}}(\{qs-\})\alpha_{n-m}(qs) \sum_{w \in (A \times S)^m} \frac{P_{\pi^{(m-1)}}(\{q^*sw\})}{P_{\pi^{(m-1)}}(\{q^*s-\})} \gamma_{n-m}(sw),$$

where the sums are taken over only those q such that qs is feasible under $\pi^{(m-1)}$ and where the last equality uses relations given by Eqs. (4.7) and (4.5). Then because of

Eqs. (4.3) and (4.4) the expression becomes

$$\left[\sum_{q\in(S\times A)^{n-m}} P_{\pi^{(m-1)}}(\{qs-\})\alpha_{n-m}(qs)\right] r_{q^*s}^{(m-1)}$$

$$\geqslant \sum_{q\in(S\times A)^{n-m}} [P_{\pi^{(m-1)}}(\{qs-\})\alpha_{n-m}(qs) r_{qs}^{(m-1)}]$$

$$= \sum_{q\in(S\times A)^{n-m}} P_{\pi^{(m-1)}}(\{qs-\})\alpha_{n-m}(qs) \sum_{w\in(A\times S)^m} \frac{P_{\pi^{(m-1)}}(\{qsw\})}{P_{\pi^{(m-1)}}(\{qs-\})}\gamma_{n-m}(sw)$$

$$= \sum_{q\in(S\times A)^{n-m}} \sum_{w\in(A\times S)^m} \alpha_{n-m}(qs)\gamma_{n-m}(sw) P_{\pi^{(m-1)}}(\{qsw\}).$$

Let

$$D_s^{n-m} = \{s_0 a_1 s_1 \cdots a_n s_n : s_{n-m} = s\}.$$

The calculation above has shown that

$$\int_{D_s^{n-m}} \alpha_{n-m}\gamma_{n-m}\, dP_{\pi^{(m)}} \geqslant \int_{D_s^{n-m}} \alpha_{n-m}\gamma_{n-m}\, dP_{\pi^{(m-1)}}. \tag{4.8}$$

(In writing the preceding integrals, we have extended the domain of α_{n-m} and γ_{n-m} to all of H_n in the obvious way.) And since β_{n-m}, considered as a function on H_n, is constant on the rightmost factors $(A \times S)^m$, we have

$$\int_{D_s^{n-m}} \beta_{n-m}\, dP_{\pi^{(m)}} \geqslant \int_{D_s^{n-m}} \beta_{n-m}\, dP_{\pi^{(m-1)}}. \tag{4.9}$$

Because of Eqs. (4.8), (4.9) and (4.2) we get

$$\int_{D_s^{n-m}} g\, dP_{\pi^{(m)}} \geqslant \int_{D_s^{n-m}} g\, dP_{\pi^{(m-1)}}. \tag{4.10}$$

Finally, take the sum over all s in S in Eq. (4.10) to get

$$\int g\, dP_{\pi^{(m)}} \geqslant \int g\, dP_{\pi^{(m-1)}} \geqslant \int g\, dP_{\pi}.$$

We proceed as far as step $n-1$ and obtain a plan $\pi^{(n-1)}$ where, by construction, each of the selection rules $\pi_2^{(n-1)}, \ldots, \pi_n^{(n-1)}$ is Markov, and where the selection rule $\pi_1^{(n-1)}$ is automatically Markov. Thus, we get the intended Markov plan $\hat{\pi}$ for Eq. (1.1) by setting $\hat{\pi} = \pi^{(n-1)}$. □

Corollary 4.2. *Under the hypotheses of the theorem, if there exists an optimal plan, then there exists an optimal Markov plan.*

The argument above is an extension and modification of the backward induction algorithm which originated with Bellman's principle of optimality (Bellman, 1957). This principle was used by Blackwell (1964) to obtain good Markov plans in finite-stage dynamic programming. The broad and thorough study by Puterman (1994) presents

several interesting examples of how the backward induction algorithm is applied in specific finite-horizon Markov decision problems.

Proof that Markov-achievability implies NFLSP. Suppose $g : H_n \to \mathbb{R}$ does not have the non-forking linear sections property. To show that g does not have the NFLSP, we will generalize the reasoning of Example 2.2. Because of Lemma 2.4 there exist $s^* \in S$, $q' = s'_0 a'_1 s'_1 \cdots a'_k$ and $q'' = s''_0 a''_1 s''_1 \cdots a''_k$ in $(S \times A)^k$, and there exist non-forking paths w^1, w^2, and w^3, in $(A \times S)^{n-k}$, where

$$w^i = a^i_{k+1} s^i_{k+1} \cdots a^i_n s^i_n \quad \text{for } i = 1,2,3$$

and a^1_{k+1}, a^2_{k+1}, and a^3_{k+1} are distinct, such that for some real number ξ with $0 < \xi \leqslant 1$,

$$\frac{\xi g(q''s^*w^1) + (1-\xi)g(q''s^*w^3) - g(q''s^*w^2)}{\xi g(q's^*w^1) + (1-\xi)g(q's^*w^3) - g(q's^*w^2)} < 0.$$

We must show that there exists a positive number ε_0, a randomized action map $\mathcal{Q}$, an initial distribution p_0, a transition law p, and a plan π in $(S, A, \mathcal{Q}, p, g)$ such that for every Markov plan $\hat{\pi}$ in $(S, A, \mathcal{Q}, p, g)$,

$$\int_{H_n} g \, dP_{\hat{\pi}} \leqslant \int_{H_n} g \, dP_{\pi} - \varepsilon_0. \tag{4.11}$$

For each s in S and each j such that $1 \leqslant j < k$, let $\mathcal{Q}(s,j)$ contain the single measure ρ such that

$$\rho(\{a'_j\}) = \tfrac{1}{2} \quad \text{and} \quad \rho(\{a''_j\}) = \tfrac{1}{2}$$

if $a'_j \neq a''_j$ (or $\rho(\{a'_j\}) = 1$ if $a'_j = a''_j$). Also, if $a = a'_j$ or $a = a''_j$ and $s'_{j+1} \neq s''_{j+1}$ let

$$p(\{s'_{j+1}\}|s'_j, j, a) = \tfrac{1}{2} \quad \text{and} \quad p(\{s''_{j+1}\}|s'_j, j, a) = \tfrac{1}{2}, \tag{4.12}$$

and if $a = a'_j$ or $a = a''_j$ and $s'_{j+1} = s''_{j+1}$ let

$$p(\{s'_{j+1}\}|s'_j, j, a) = 1. \tag{4.13}$$

For j such that $k < j \leqslant n$, let $\mathcal{Q}(s,j)$ contain all those probability measures which give unit mass to either $\{a^1_j\}$ or $\{a^2_j\}$ or $\{a^3_j\}$.

Further, for each a in A, let

$$p(\{s^*\}|s'_{k-1}, k-1, a) = p(\{s^*\}|s''_{k-1}, k-1, a) = 1,$$

let

$$\mathfrak{H} = \{s_0 a_1 s_1 \cdots s_{k-1} a_k : s_j \in \{s'_j, s''_j\} (0 \leqslant j \leqslant k-1) \text{ and } a_j \in \{a'_j, a''_j\} (1 \leqslant j \leqslant k)\},$$

and suppose that the initial distribution p_0 satisfies

$$p_0(\{s'_0\}) = p_0(\{s''_0\}) = \tfrac{1}{2} \tag{4.14}$$

(or $p_0(\{s'_0\}) = 1$ if $s'_0 = s''_0$).

Let $\mathcal{Q}(s^*,k) = \{\mu, \nu\}$, where

$$\mu(\{a_{k+1}^1\}) = \xi \quad \text{and} \quad \mu(\{a_{k+1}^3\}) = 1 - \xi \quad \text{and} \quad \mu(\{a_{k+1}^2\}) = 1.$$

To create the plan π, observe that Eqs. (4.12)–(4.14) allow us to define the selection rules $\pi_1, \pi_2, \ldots, \pi_{k-1}$ so that each is Markov and so that for each q in $\mathfrak{H}$, the P_π-probability of the set $\{qs^*-\}$ is the same. We define the non-Markov selection rule π_k so that for q in $\mathfrak{H}$,

$$\pi_k(qs^*) = \begin{cases} \mu & \text{if} \quad \xi g(qsw^1) + (1-\xi)g(qsw^3) \geqslant g(qsw^2), \\ \nu & \text{if} \quad \xi g(qsw^1) + (1-\xi)g(qsw^3) < g(qsw^2). \end{cases}$$

To finish the definition of π, we define $\pi_{k+1}, \pi_{k+2}, \ldots, \pi_n$ in such a way that for $i = 1,2,3$, respectively, after action a_{k+1}^i is used at time k, the plan π follows the single history w^i from that time onward. Thus, for fixed i, each of the elements of $\{qs^*w^i : q \in \mathfrak{H}\}$ receives equal probability under P_π. Notice that the "non-forking" nature of w^1, w^2, and w^3 makes it possible to arrange the Markov selection rules $\pi_j\,(k < j \leqslant n)$ in this way. If two paths w^i and w^j had split apart after moving in tandem for a while, then the definition of the selection rules would have been ambiguous. Let M be the number of (distinct) elements in the set $\mathfrak{H}$. Then

$$\int g\,\mathrm{d}P_\pi = \frac{1}{M}\left[\sum_{\{q:\pi_k(qs^*)=\mu\}} [\xi g(qs^*w^1) + (1-\xi)g(qsw^3)] + \sum_{\{q:\pi_k(qs^*)=\nu\}} g(qs^*w^2)\right]. \tag{4.15}$$

Now, let $\hat{\pi}$ be a Markov plan in $(S, A, \mathcal{Q}, p, g)$. It is only when $j = k$ and $s_j = s^*$ that the transition probability $p(\cdot|s_j, j, a)$ depends on the action a. Thus, it is only when $j = k$ and $s_j = s^*$ that the definition of $\hat{\pi}_j$ is an issue. Furthermore, since $\hat{\pi}$ is Markov, either we have

$$\hat{\pi}_k(qs^*) = \nu \qquad \text{for all } q \text{ in } (S \times A)^{k-1} \tag{4.16}$$

or

$$\hat{\pi}_k(qs^*) = \mu \qquad \text{for all } q \text{ in } (S \times A)^{k-1}. \tag{4.17}$$

Now, because of Eqs. (4.16), (4.17) and (4.15) we conclude that there exists $\varepsilon_0 > 0$ which satisfies the desired inequality (4.11). □

5. Functions having the general linear sections property

The main purpose of this section is to demonstrate the next theorem, which gives an exact characterization of those payoff functions which have the GLSP. It follows from Proposition 3.3 and Theorem 3.7 that all such functions are Markov-achievable.

Theorem 5.1. *The payoff function* $g : H_n \to \mathbb{R}$ *has the GLSP if and only if there exist mappings*

$$u_i : S \times A \times S \to \mathbb{R}, \qquad i = 1, \ldots, n$$

and

$$v_j : S \times A \times S \to [0, \infty), \qquad j = 1, \ldots, n-1$$

such that for each element $s_0 a_1 s_1 \cdots a_n s_n$ *of* H_n,

$$g(s_0 a_1 s_1 \cdots a_n s_n) = u_1(s_0 a_1 s_1) + \sum_{i=2}^{n} u_i(s_{i-1} a_i s_i) \prod_{j=1}^{i-1} v_j(s_{j-1} a_j s_j). \tag{5.1}$$

Of course, the condition in the theorem above includes the standard additive utility payoff of classical dynamic programming (see Example 2.6). It also includes the usual finite-stage multiplicative utility models. Multiplicative utilities were used by Howard and Matheson (1972) and by Rothblum (1984). Payoff functions which arise from neither additive nor multiplicative utilities, or their limits, are sometimes referred to as "non-standard". Several such non-standard optimality criteria are discussed by White (1993). Another different, but yet natural, criterion is used by Feinberg and Shwartz (1995).

The following lemma is analogous to Lemma 3.9 and formulates the general linear sections property (Definition 2.5) in terms of slope and y-intercept:

Lemma 5.2. *The payoff function* $g : H_n \to \mathbb{R}$ *has the GLSP if and only if for each* k $(1 \leqslant k \leqslant n-1)$, *there exist bounded mappings* $\alpha_k : H_k \to [0, \infty)$, $\beta_k : H_k \to \mathbb{R}$, *and* $\gamma_k : H_{n-k} \to \mathbb{R}$ *such that for each* q *in* $(S \times A)^k$, *each* s *in* S, *and each* w *in* $(A \times S)^{n-k}$,

$$g(qsw) = \alpha_k(qs)\gamma_k(sw) + \beta_k(qs). \tag{5.2}$$

Proof. Immediate from Definitions 2.3 and 2.5. □

Proof of Theorem 5.1. Suppose g has the GLSP. Then by Lemma 5.2, for each k $(1 \leqslant k \leqslant n-1)$ there exist functions $\alpha_k : H_k \to [0, \infty)$, $\beta_k : H_k \to \mathbb{R}$, $\gamma_k : H_{n-k} \to \mathbb{R}$ (for each $s_k \in S$, $\gamma_k(s_k -)$ is not constant) such that for each $s_0 a_1 s_1 \cdots a_n s_n \in H_n$,

$$g(s_0 a_1 \cdots s_n) = \alpha_k(s_0 a_1 \cdots s_k)\gamma_k(s_k a_{k+1} \cdots s_n) + \beta_k(s_0 a_1 \cdots s_k). \tag{5.3}$$

Define u_1, v_1, and u_n on H_1 by

$$u_1 = \beta_1, \quad v_1 = \alpha_1, \quad \text{and} \quad u_n = \gamma_{n-1}. \tag{5.4}$$

Now, suppose m satisfies $2 \leqslant m \leqslant n-1$, and let s be an arbitrary element of S. If for each q in $(S \times A)^{m-1}$, $\alpha_{m-1}(qs) = 0$, then for each a in A and each y in S, let

$$u_m(say) = v_m(say) = 1;$$

if there exists q^* in $(S \times A)^{m-1}$ such that $\alpha_{m-1}(q^* s) \neq 0$, let

$$u_m(say) = \frac{\beta_m(q^* say) - \beta_{m-1}(q^* s)}{\alpha_{m-1}(q^* s)} \tag{5.5}$$

and

$$v_m(say) = \frac{\alpha_m(q^*say)}{\alpha_{m-1}(q^*s)}. \tag{5.6}$$

We must show that these definitions of $u_m : H_1 \to \mathbb{R}$ and $v_m : H_1 \to [0,\infty)$ do not depend on q^*. To do so, suppose $q' \in (S \times A)^{m-1}$ and $q'' \in (S \times A)^{m-1}$ with $\alpha_{m-1}(q's) \neq 0$ and $\alpha_{m-1}(q''s) \neq 0$. Now, let

$$\triangle_1 = \frac{\alpha_m(q'say)}{\alpha_{m-1}(q's)},$$

$$\triangle_2 = \frac{\alpha_m(q''say)}{\alpha_{m-1}(q''s)},$$

$$\triangledown_1 = \frac{\beta_m(q'say) - \beta_{m-1}(q's)}{\alpha_{m-1}(q's)},$$

and

$$\triangledown_2 = \frac{\beta_m(q''say) - \beta_{m-1}(q''s)}{\alpha_{m-1}(q''s)};$$

We will show that $\triangle_1 = \triangle_2$ and $\triangledown_1 = \triangledown_2$.

If $q \in (S \times A)^{m-1}$ and $w \in (A \times S)^{n-m}$, Eq. (5.3) gives

$$g(qsayw) = \alpha_{m-1}(qs)\gamma_{m-1}(sayw) + \beta_{m-1}(qs)$$

and

$$g(qsayw) = \alpha_m(qsay)\gamma_m(yw) + \beta_m(qsay).$$

Thus for $q = q'$ (with $i = 1$) and for $q = q''$ (with $i = 2$), we have

$$\gamma_{m-1}(sayw) = \triangle_i\gamma_m(yw) + \triangledown_i. \tag{5.7}$$

By Eq. (5.7),

$$0 = (\triangle_1 - \triangle_2)\gamma_m(yw) + (\triangledown_1 - \triangledown_2).$$

Since w is arbitrary, and since $\gamma_m(y-)$ is not constant, we must have $\triangle_1 = \triangle_2$ and $\triangledown_1 = \triangledown_2$, as desired. Therefore, the functions u_m and v_m are well defined.

To begin to prove Eq. (5.1), let

$$\Phi = \{k\colon 1 \leqslant k \leqslant n-1, \alpha_k(s_0a_1\cdots s_k) = 0\}.$$

Assume temporarily that Φ is non-empty, and let M be its smallest element. Then because of Eq. (5.3) we get

$$g(s_0a_1\cdots s_n) = \beta_M(s_0a_1\cdots s_M) \tag{5.8}$$

and because of Eq. (5.6), for $1 \leqslant k \leqslant M-1$, we have

$$\prod_{j=1}^{k} v_j(s_{j-1}a_js_j) = \alpha_k(s_0a_1\cdots s_k) \neq 0, \tag{5.9}$$

and

$$v_M(s_{M-1}a_Ms_M) = 0. \tag{5.10}$$

Then from Eqs. (5.4), (5.5), (5.9), and (5.10), it follows that

$$\begin{aligned} &u_1(s_0a_1s_1) + \sum_{i=2}^{n} u_i(s_{i-1}a_is_i)\prod_{j=1}^{i-1} v_j(s_{j-1}a_js_j) \\ &= \beta_1(s_0a_1s_1) + \sum_{i=2}^{M}[\beta_i(s_0a_1\cdots s_i) - \beta_{i-1}(s_0a_1\cdots s_{i-1})] \\ &= \beta_M(s_0a_1\cdots s_M). \end{aligned}$$

By Eq. (5.8), these expressions are equivalent to $g(s_0a_1\cdots s_n)$, and thus Eq. (5.1) holds when Φ is non-empty.

On the other hand, if Φ is empty, then Eq. (5.9) holds for each k $(1\leqslant k\leqslant n-1)$. Therefore,

$$\begin{aligned} &u_1(s_0a_1s_1) + \sum_{i=2}^{n} u_i(s_{i-1}a_is_i)\prod_{j=1}^{i-1} v_j(s_{j-1}a_js_j) \\ &= u_1(s_0a_1s_1) + u_n(s_{n-1}a_ns_n)\prod_{j=1}^{n-1} v_j(s_{j-1}a_js_j) + \sum_{i=2}^{n-1} u_i(s_{i-1}a_is_i)\prod_{j=1}^{i-1} v_j(s_{j-1}a_js_j) \\ &= \beta_1(s_0a_1s_1) + \gamma_{n-1}(s_{n-1}a_ns_n)\alpha_{n-1}(s_0a_1\cdots s_{n-1}) \\ &\quad + \sum_{i=2}^{n-1}[\beta_i(s_0a_1\cdots s_i) - \beta_{i-1}(s_0a_1\cdots s_{i-1})] \\ &= \alpha_{n-1}(s_0a_1\cdots s_{n-1})\gamma_{n-1}(s_{n-1}a_ns_n) + \beta_{n-1}(s_0a_1\cdots s_{n-1}) \\ &= g(s_0a_1\cdots s_n). \end{aligned}$$

This completes the proof of Eq. (5.1).

Conversely, if for each $s_0a_1\cdots s_n$ in H_n, Eq. (5.1) holds, then for $1\leqslant k\leqslant n-1$, let

$$\alpha_k(s_0a_1\cdots s_k) = \prod_{j=1}^{k} v_j(s_{j-1}a_js_j),$$

let

$$\beta_1 = u_1,$$

and for $2\leqslant k\leqslant n-1$, let

$$\beta_k(s_0a_1\cdots s_k) = u_1(s_0a_1s_1) + \sum_{i=2}^{k} u_i(s_{i-1}a_is_i)\prod_{j=1}^{i-1} v_j(s_{j-1}a_js_j);$$

let

$$\gamma_{n-1} = u_n$$

and for $1 \leqslant k \leqslant n-2$, let

$$\gamma_k(s_k a_{k+1} \cdots s_n) = u_{k+1}(s_0 a_1 s_1) + \sum_{i=k+2}^{n} u_i(s_{i-1} a_i s_i) \prod_{j=1}^{i-1} v_j(s_{j-1} a_j s_j).$$

Then for each k $(1 \leqslant k \leqslant n-1)$,

$$g(s_0 a_1 \cdots s_n) = \alpha_k(s_0 a_1 \cdots s_k)\gamma_k(s_k a_{k+1} \cdots s_n) + \beta_k(s_0 a_1 \cdots s_k).$$

Therefore, by Lemma 5.2, g has the GLSP. □

Acknowledgements

We are grateful to the editor and to an anonymous referee for their careful reading of the manuscript and for their valuable suggestions.

References

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton NJ.

Blackwell, D., 1964. Memoryless strategies in finite-stage dynamic programming. Ann. Math. Statist. 35, 863–865.

Blackwell, D., 1965. Discounted dynamic programming. Ann. Math. Statist. 36 (1965) 226–235.

Dubins, L., Savage, L., 1976. Inequalities for Stochastic Processes (How to Gamble if You Must). Dover, New York.

Dynkin, E., Yushkevich, A., 1979. Controlled Markov Processes. Springer, Berlin.

Feinberg, E., 1982. Controlled Markov processes with arbitrary numerical criteria. Theory Probab. Appl. 27, 486–503.

Feinberg, E., Shwartz, A., 1995. Constrained Markov decision models with weighted discounted rewards. Math. Oper. Res. 20, 302–320.

Hill, T., Pestien, V., 1987. The existence of good Markov strategies for decision processes with general payoffs. Stochastic Process. Appl. 24, 61–76.

Howard, R.A., Matheson, J.E., 1972. Risk sensitive Markov decision processes. Management Sci. 18, 356–369.

Hinderer, K., 1970. Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Springer, Berlin.

Larson, R., Casti, J., 1978. Principles of Dynamic Programming, Part I. Marcel Dekker, New York.

Maitra, A., Sudderth, W., 1996. Discrete Gambling and Stochastic Games. Springer, New York.

Pestien, V., Wang, X., 1993. Finite-stage reward functions having the Markov adequacy property. Stochastic Process. Appl. 46, 129–151.

Puterman, Martin L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York.

Rothblum, U., 1984. Multiplicative Markov decision chains. Math. Oper. Res. 9, 6–24.

Schäl, M., Sudderth, W., 1987. Stationary policies and Markov policies in dynamic programming. Probab. Theory Rel. Fields 74, 91–111.

White, D., 1993. Markov Decision Processes. Wiley, Chichester, UK.